

Quadratic equations over free groups are NP-complete

O. Kharlampovich, I.G. Lysënok, A.G. Myasnikov, N.W.M. Touikan

February 26, 2008

Abstract

We prove that the problems of deciding whether a quadratic equation over a free group has a solution is NP-complete

1 Introduction

The study of quadratic equations over free groups probably started with [Mal62] and has been deepened extensively ever since. One of the reasons research in this topic has been so fruitful is the deep connection between quadratic equations and the topology of surfaces.

In this paper we will show that the problem of deciding if a quadratic equation over a free group is NP-complete. This problem was shown to be decidable in [CE81]. In addition it was shown in [Ol'89], [GK92], and [GL92] that if n , the number of variables, is fixed, then deciding if a quadratic equation has a solution can be done in time polynomial in the sum of the lengths of the coefficients. These results imply that the problem is solvable in at most exponential time. We will improve on this by proving (see Theorem 2.1) that deciding if an quadratic equation over a free group has a solution is in NP.

In [DR99] it is shown that deciding if a quadratic word equation has a solution is NP-hard. We will prove (see Theorem 3.10) that deciding if a quadratic equations over a free group has a solution is also NP-hard. Our proofs are geometric, relying on the topological results of [Ol'89] and disc diagram techniques.

2 Quadratic equations over free groups are in NP

For a finite alphabet S we denote by S^* the free monoid with involution with basis S and for $w \in S^*$, we denote by w^{-1} its involution. We denote by $F(S)$ the free group on S .

2.1 Standard form

A quadratic equation E with variables $\{x_i, y_i, z_j\}$ and coefficients $\{w_i, d\} \in F(A)$ is said to be in *standard form* if its coefficients are expressed as freely and cyclically reduced words in A^* and E has either the form:

$$\prod_{i=1}^g [x_i, y_i] \prod_{j=1}^{m-1} z_j^{-1} w_j z_j d = 1 \text{ or } \prod_{i=1}^g [x_i, y_i] = 1 \quad (1)$$

where $[x, y] = x^{-1}y^{-1}xy$, in which case we say it is *orientable* or it has the form

$$\prod_{i=1}^g x_i^2 \prod_{j=1}^{m-1} z_j^{-1} w_j z_j d = 1 \text{ or } \prod_{i=1}^g x_i^2 d = 1 \quad (2)$$

in which case we say it is non orientable. The *genus* of a quadratic equation is the number g in (1) and (2) and m is the number of coefficients. If $g = 0$ then we will define E to be orientable. If E is a quadratic equation we define its *reduced euler characteristic*, $\overline{\chi}$ as follows:

$$\overline{\chi}(E) = \begin{cases} 2 - 2g & \text{if } E \text{ is orientable} \\ 2 - g & \text{if } E \text{ is not orientable} \end{cases}$$

It is a well known fact that an arbitrary quadratic equation over a free group can be brought to a standard form in time polynomial in its length.

2.2 Ol'shanskii's result

In sections 2.3 and 2.4 of [Ol'89] it is shown that a quadratic equation E in standard form has a solution if and only if for some $n \leq 3(m - \overline{\chi}(E))$,

- (i) there is a set $P = \{p_1, \dots, p_n\}$ of variables and a collection of m discs D_1, \dots, D_m such that,
- (ii) the boundaries of these discs are directed labeled graphs such that each edge has a label in P and each $p_j \in P$ occurs exactly twice in the union of boundaries;
- (iii) if we glue the discs together by edges with the same label, respecting the edge orientations, then we will have a collection $\Sigma_0, \dots, \Sigma_l$ of closed surfaces and the following inequalities: if E is orientable then each Σ_i is orientable and

$$\left(\sum_{i=0}^l \chi(\Sigma_i) \right) - 2l \geq \overline{\chi}(E)$$

if E is non-orientable either at least one Σ_i is non-orientable and

$$\left(\sum_{i=0}^l \chi(\Sigma_i) \right) - 2l \geq \overline{\chi}(E)$$

or, each Σ_i is orientable and

$$\left(\sum_{i=0}^l \chi(\Sigma_i) \right) - 2l \geq \overline{\chi}(E) + 2$$

and

- (iv) there is a mapping $P \rightarrow A^*$ such that upon substitution, the coefficients w_1, \dots, w_{m-1} and d can be read without cancellations around the boundaries of D_1, \dots, D_{m-1} and D_m , respectively; and finally that
- (v) if E is orientable the discs D_1, \dots, D_m can be oriented so that w_i is read clockwise around ∂D_i and d is read clockwise around ∂D_m , moreover all these orientations must be compatible with the glueings.

We note that the bounds in (iii) are not given explicitly in that paper, but they follow immediately from the discussion on cutting up diagrams into so-called simple diagrams, see [OI'89] for details.

2.3 The certificate

The result of section 2.2 enables us to construct a good certificate.

Theorem 2.1. *For a quadratic equation E in standard form there is a certificate of size bounded by $2(|w_1| + \dots + |w_m| + |d| + 3(2g + m))$ that can be checked in polynomial time.*

Proof. The certificate will consist of the following:

1. A collection of variables $P = \{p_1, \dots, p_n\}$.
2. A collection of substitutions $\bar{\psi} = \{p_i \mapsto a_i, i = 1 \dots n\}$ where $a_i \in A^*$ and $n < 3(2g + m)$.
3. A collection of words in P^*

$$\mathcal{C} = \begin{cases} C_1 = p_{11}^{\epsilon_{11}} \dots p_{11}^{\epsilon_{1j_1}} \\ \dots \\ C_m = p_{m1}^{\epsilon_{m1}} \dots p_{mj_m}^{\epsilon_{mj_m}} \end{cases}$$

with $p_{ij} \in P, \epsilon_{ij} \in \{-1, 1\}$ and each $p_i \in P$ occuring exactly twice.

The C_i 's represent the labels of the boundaries of the discs D_1, \dots, D_m so checking conditions (i) and (ii) of Section 2.2 can be done quickly, moreover we see that the size of \mathcal{C} is at most $2n \leq 6(2g + m)$.

$\bar{\psi}$ extends to a monoid homomorphism $\psi : P^* \rightarrow A^*$. (iv) can also be verified quickly since for $i = 1, \dots, m-1$ we just need to check that some cyclic permutation of $\psi(C_i)$ is equal to w_i and some cyclic permutation of $\psi(C_m)$ is equal to d . Moreover, since the equality is graphical we have that

$$|a_1| + \dots + |a_n| \leq |w_1| + \dots + |w_m| + |d|$$

Therefore the size of the certificate is bounded as advertised. All that is left is to determine the topology of the glued together discs. We describe the algorithm without too much detail.

Step 1: Built a forest of discs: We make a graph Γ such that each vertex $v_i \in V(\Gamma)$ corresponds the disc D_i and each edge $e_j \in E(\Gamma)$ corresponds to the variable $p_j \in P$. The edge e_k goes from v_i to v_j if and only if the variable p_k occurs in the boundary of D_i and in the boundary of D_j or if $i = j$ then there are two different occurrences of the variable p_k . We construct a spanning forest \mathcal{F} . This enables us to count the number of connected components $\Sigma_0, \dots, \Sigma_l$.

Step 2: Determine orientability: For each maximal tree $T_r \subset \mathcal{F}$ we get a “tree of discs” by glueing together only the pairs of edges whose labels correspond to elements of $E(T_r)$. The resulting tree of discs is a simply connected topological space that can be embedded in the plane and we can read a cyclic word $c(T_r)$ in P^* along its boundary. The surface Σ_r obtained by glueing together the remaining paired edges of the tree of discs will be orientable only if whenever $p_j^{\pm 1}$ occurs in $c(T_r)$ then $p_j^{\mp 1}$ also occurs. We can also check (v) at this point.

Step 3: Compute Euler characteristic: The identification of the boundary of the discs with graphs, enables us to think of the discs as polygons. If a disc D_i has N_i sides then we give each corner of D_i an angle of $\pi(N_i - 2)/N_i$. Then for each tree of discs produced in the previous step, we identify the remaining pairs of edges to get the surfaces $\Sigma_0, \dots, \Sigma_l$, which now have an extra angular structure. To each Σ_i , we can apply the Combinatorial Gauss-Bonnet Theorem which states that for an angled two-complex X ,

$$2\pi\chi(X) = \sum_{f \in X^{(2)}} \kappa(f) + \sum_{v \in X^{(0)}} \kappa(v)$$

where $X^{(2)}$ is the set of faces and $X^{(0)}$ is the set of vertices. This angle assingment gives each face f a curvature $\kappa(f) = 0$ and each vertex has curvature

$$\kappa(v) = 2\pi - \left(\sum_{c \in \text{link}(v)} \angle(c) \right)$$

i.e. $\kappa(v)$ is 2π minus the sum of the angles that meet at v .

With an appropriate data structure one can perform steps 1-3 (not necessarily in sequential order) in at most quadratic time in the size of \mathcal{C} . Once all that is done, verifying the inequalities of (iii) is easy and we are finished. \square

3 Quadratic equations over free groups are NP-hard

We will present the bin packing problem which is known to be NP-complete and show that it is equivalent deciding if a certain type of quadratic equation has a solution.

3.1 Bin Packing

Problem 3.1 (Bin Packing).

- *INPUT:* A k -tuple of positive integers (r_1, \dots, r_k) and positive integers B, N .
- *QUESTION:* Is there a partition of $\{1, \dots, k\}$ into N subsets

$$\{1, \dots, k\} = S_1 \sqcup \dots \sqcup S_N$$

such that for each $i = 1, \dots, N$ we have

$$\sum_{j \in S_i} r_j \leq B \tag{3}$$

This problem is NP-hard in the strong sense (see [GJ79] p.226), i.e. there are NP-hard instances of this problem when both B and the r_j are bounded by a polynomial function of k .

Let $t = NB - \sum_{i=1}^k r_i$. Then by replacing (r_1, \dots, r_k) by the $k+t$ -tuple $(r_1, \dots, r_k, \dots, 1, \dots, 1)$ we can assume that the inequalities (3) are actually equalities. This modified version is still NP hard in the strong sense. We state it explicitly:

Problem 3.2 (Exact Bin Packing).

- *INPUT: A k -tuple of positive integers (r_1, \dots, r_k) and positive integers B, N .*
- *QUESTION: Is there a partition of $\{1, \dots, k\}$ into N subsets*

$$\{1, \dots, k\} = S_1 \sqcup \dots \sqcup S_N$$

such that for each $i = 1, \dots, N$ we have

$$\sum_{j \in S_i} r_j = B \tag{4}$$

The authors warmly thank Laszlo Babai for drawing their attention to this problem in connection to tiling problems.

3.2 Tiling discs

Throughout this section we will consider the discs to be embedded in \mathbb{E}^2 and will always read clockwise around closed curves.

Definition 3.3. A $[a, b^n]$ -disc is a disc as in section 2.2 along whose boundary one can read the cyclic word $[a, b^n]$.

Definition 3.4. A $[a, b^n]$ -ribbon is a rectangular cell complex obtained by attaching $[a, b^j]$ -discs by their a -labeled edges, such that we can read $[a, b^n]$ along its boundary. The *top* of an $[a, b^n]$ ribbon is the boundary subpath along which we can read the word b^{-n} , the *bottom* is the boundary subpath along which we can read the word b^n .

Definition 3.5. Let D be a disc tiled by $[a, b^n]$ -discs, we define the a -pattern of D to be a graph defined as follows:

1. In the middle of each a -labeled edge put a vertex.
2. Between any two vertices contained in the same $[a, b^n]$ -disc draw an edge.

Connected components of a -patterns are called a -tracks

Lemma 3.6. A disc D tiled by finitely many $[a, b^n]$ -discs cannot have any circular a -tracks.

Proof. It is clear that every a -track is a graph whose vertices have valency at most 2. If an a -track t has vertices of valency 1 then they must lie on ∂D .

Suppose towards a contradiction that D has a circular a -track c . Then c divides D into two components: an interior and an exterior. If we examine the interior we see that it is a planar union of discs with only the letter b

occurring on its boundary, it follows that the interior contains a disc D' with circular a -track. Repeating the argument we find that D must have infinitely many cells which is a contradiction. \square

Corollary 3.7. *If D is a disc tiled by finitely many $[a, b^n]$ -discs, then the cyclic word read around ∂D cannot contain only the letter b .*

Corollary 3.8. *We cannot tile a sphere with finitely many coherently oriented $[a, b^n]$ -discs.*

Proposition 3.9. *Suppose that D is a disc with boundary label $[a^N, b^B]$ that is covered by $[a, b^n]$ -discs, then it is obtained from a collection of M $[a, b^B]$ -ribbons R_1, \dots, R_M such that the bottom of R_{i+1} is glued to the top of R_i , $i = 1, \dots, M$.*

Proof. We proceed by induction on N . If $N = 1$, then we consider the a -track t starting at one of the edges of ∂D labeled a . t must touch the other edge labeled a in ∂D . Let $R(t)$ be the subset of D consisting of the $[a, b^n]$ -discs that t intersects. We note that $R(t)$ can be obtained by making some identifications in the top and bottom of some $[a, b^B]$ -ribbon, but $R(t) \subset D$, which means on one hand that if $R(t)$ is not simply connected then some subset of $\partial R(t)$ is a circle that bounds a disc inside D , this disc can only have b 's in its label contradicting Corollary 3.7. It follows that $R(t)$ is a ribbon and it contains every a -labeled edge in D , so we must have $R(t) = D$.

Suppose the hypothesis held for all $L \leq N - 1$ and suppose that we could read $[a^N, b^B]$ along ∂D . We divide ∂D into four arcs l_a, t_b, r_a, b_b that have labels a^{-N}, b^{-B}, a^N, b^B respectively, i.e. the left, top, right and bottom sides. Let e be the edge with label a that touches the vertex between l_a and t_b . Let t be the corresponding a -track. Let $R(t)$ be as above, since $D \subset \mathbb{E}^2$ it is easy to see that t cannot be a line from l_a to l_a , therefore t must go from l_a to some edge e' in r_a .

Suppose towards a contradiction that e' was not the edge in r_a that touched the vertex v between t_b and r_a . Let f be the edge in r_a that touches v , and let u be the corresponding a -track, since a -tracks cannot cross we have that u must also end in r_a which is a contradiction.

By the same argument as in the case $N = 1$ we have that $R(t)$ must be an embedded ribbon. By Corollary 3.7 we must have that t_b is contained in the top of $R(t)$, which means that $R(t)$ is an embedded $[a, b^B]$ -ribbon and if we remove $R(t)$ from D , then what remains is a disc D' such that we can read $[a^{N-1}, b^B]$ along the boundary. So by induction the result follows. \square

3.3 A special genus zero quadratic equation

Equipped with Proposition 3.9 we shall deduce NP hardness of the following equation:

$$\prod_{j=1}^k z_j^{-1} [a, b^{n_j}] z_j = [a^N, b^B] \quad (5)$$

By the results in section 2.2, (5) has a solution if and only if there is a collection of discs D_j with boundary labels $[a, b^{n_j}]$ for $j = 1 \dots k$ respectively and a disc D_m with boundary label $[a^N, b^B]$ such that, glued together in a

way that respect labels and orientation of edges, form a union of spheres (this is forced by the first inequality in (iii), section 2.2).

Theorem 3.10. *Deciding if the quadratic equation (5) with coefficients*

$$[a, b^{n_1}], \dots, [a, b^{n_k}] \text{ and } [a^N, b^B]$$

has a solution is equivalent to deciding if problem 3.2; with input (n_1, \dots, n_m) and positive integers B, N ; has a positive answer.

Proof. “Bin packing \Rightarrow solution.” Suppose that Problem 3.2 has a positive answer on the specified inputs. For each subset S_i of the given partition of $\{1, \dots, k\}$ we form a $[a, b^B]$ -ribbon R_i by glueing together the $[a, b^{n_j}]$ -discs for $j \in S_i$, this is possible by (iv) in section 2.2 and equation (4). We then construct one hemisphere by glueing the ribbons R_1, \dots, R_N . The other hemisphere is the remaining disc with boundary label $[a^N, b^B]^{-1}$, the resulting sphere proves the solvability of (5) with the given coefficients.

“Solution \Rightarrow bin packing.” If (5) has a solution then there is a union of spheres tiled with $[a, b^{n_i}]$ -discs and one $[a^N, b^B]^{-1}$ -disc, moreover these discs are coherently oriented. By condition (v) and Corollary 3.8 there can only be one sphere: the sphere S_0 containing the unique $[a^N, b^B]^{-1}$ -disc. If we remove this $[a^N, b^B]^{-1}$ -disc from S_0 what remains will be a disc D with boundary label $[a^N, b^B]$ tiled with $[a, b^{n_i}]$ -discs. Applying Proposition 3.9 divides D into ribbons R_1, \dots, R_N and we immediately see that these ribbons provide a partition of $\{n_1, \dots, n_k\}$, showing that Problem 3.2 has a positive solution on the given input. \square

References

- [CE81] Leo P. Comerford, Jr. and Charles C. Edmunds. Quadratic equations over free groups and free products. *J. Algebra*, 68(2):276–297, 1981.
- [DR99] Volker Diekert and John Michael Robson. Quadratic word equations. In *Jewels are forever*, pages 314–326. Springer, Berlin, 1999.
- [GJ79] Michael R. Garey and David S. Johnson. *Computers and intractability*. W. H. Freeman and Co., San Francisco, Calif., 1979. A guide to the theory of NP-completeness, A Series of Books in the Mathematical Sciences.
- [GK92] R. I. Grigorchuk and P. F. Kurchanov. On quadratic equations in free groups. In *Proceedings of the International Conference on Algebra, Part 1 (Novosibirsk, 1989)*, volume 131 of *Contemp. Math.*, pages 159–171, Providence, RI, 1992. Amer. Math. Soc.
- [GL92] R. I. Grigorchuk and I. G. Lysionok. A description of solutions of quadratic equations in hyperbolic groups. *Internat. J. Algebra Comput.*, 2(3):237–274, 1992.
- [Mal62] A. I. Mal’cev. On the equation $zxyx^{-1}y^{-1}z^{-1} = aba^{-1}b^{-1}$ in a free group. *Algebra i Logika Sem.*, 1(5):45–50, 1962.
- [Ol’89] A. Yu. Ol’shanskiĭ. Diagrams of homomorphisms of surface groups. *Sibirsk. Mat. Zh.*, 30(6):150–171, 1989.